Ctrl-P

Ctrl-P: Temporal Control of Prosodic Variation for Speech Synthesis

Devang S Ram Mohan[1], Vivian Hu[1*], Tian Huey Teh[1*], Alexandra Torresquintero[1], Christopher G. R. Wallis[1], Marlene Staib[1], Lorenzo Foglianti[1], Jiameng Gao[1], Simon King[1,2]*

[1]Papercup Technologies Ltd., [2]University of Edinburgh

{devang,vivian,tian}@papercup.com

Abstract

Text does not fully specify the spoken form, so text-to-speech models must be able to learn from speech data that vary in ways not explained by the corresponding text. One way to reduce the amount of unexplained variation in training data is to provide acoustic information as an additional learning signal. When generating speech, modifying this acoustic information enables multiple distinct renditions of a text to be produced.

Since much of the unexplained variation is in the prosody, we propose a model that generates speech explicitly conditioned on the three primary acoustic correlates of prosody: F_0, energy and duration. The model is flexible about how the values of these features are specified: they can be externally provided, or predicted from text, or predicted then subsequently modified.

Compared to a model that employs a variational auto-encoder to learn unsupervised latent features, our model provides more interpretable, temporally-precise, and disentangled control. When automatically predicting the acoustic features from text, it generates speech that is more natural than that from a Tacotron 2 model with reference encoder. Subsequent human-in-the-loop modification of the predicted acoustic features can significantly further increase naturalness.

Index Terms: text-to-speech, controllable speech synthesis

1. Introduction

There are generally multiple ways in which a given text can be spoken. These distinct renditions may be the result of semantic distinctions, or different speaking styles, or simply natural random variation. In all cases, the differences are acoustic and are not fully specified by the text. Treating this variation as unwanted noise, and averaging it away, results in a lack of variation in the synthesised speech [1]. 'Average prosody' is probably meaningless and not the same as 'default' prosody [2].

A popular approach to handling this unexplained variability is to learn a latent space [3, 4, 5, 6]. During inference, a sampled embedding from this latent space provides the information missing from the text. However, this approach can lead to undesirable artefacts in the synthetic speech [7, 8] and, since the latent space is learned in an unsupervised fashion, it is difficult to choose an appropriate embedding to convey a specific desired rendition of the text.

An alternative is to control specific acoustic features that correlate with prosodic variation. Since these can be automatically estimated from speech, it is simple to annotate the training data with their values. These acoustic features offer a direct way to synthesise prosodically distinct renditions of a text.

In our proposed model, we use three acoustic correlates of prosodic variation: F_0, energy and duration [9, 10]. Their values are specified per-phone in the force-aligned reference speech during training. Our modified version of the Tacotron 2 encoder-decoder model [11] attends over a concatenation of the encoder outputs and these acoustic features. The supervised nature of these features ensures stability of model training.

We also add to the model an acoustic feature predictor (AFP) to predict per-phone acoustic feature values, given the encoder outputs [12, 13]. This enables the model to produce natural synthesised speech from text alone, without requiring any additional inputs, whilst offering the option of control when desired.

The model thus contains interpretable, disentangled acoustic features that can be controlled at any desired temporal granularity, from individual phones to the whole utterance. This allows external human-in-the-loop control to generate multiple prosodically-distinct renditions of a given text.[1] Since the model predicts reasonable 'default' values of these features from text, external control only needs to specify the subset of values to be changed.

2. Related work

Modelling prosodic variation has been an area of research interest for decades. Unit-selection approaches include explicitly capturing variation in the recorded speech database [14]. Statistical parametric synthesisers used regression trees to map paralinguistic features to acoustic model parameters [15, 16]. Given the ability of neural approaches to generate much more natural speech than these older systems [11, 17, 18], recent research has focused on different ways of modelling prosodic variation within these approaches.

The model proposed is distinguished from preceding work by providing control over explicit acoustic features with good temporal precision. By 'temporal precision' we mean both the ability to perform control at specified locations (e.g., per phone) and for those changes to result in localised changes in the generated speech.

2.1. Explicit acoustic features vs. learnt latent dimensions

A popular approach to incorporate prosodic variation into neural TTS systems is to treat it as a residual component [5] – that is, as acoustic variation not predictable from text – and to learn a latent space which captures this information [3, 4, 5, 6, 19, 20]. Most commonly, the latent space is an embedding of an acoustic reference mel spectrogram, e.g., [3, 5, 19]. Learnt latent spaces are generally uninterpretable and the dimensions are typically entangled. This renders the latent dimensions inconvenient for use as external control 'levers'. Model estimation procedures to encourage disentanglement have been proposed [19], but these appear to make the model highly sensitive to hyperparameter

[*]These authors contributed equally to this work

[1]Samples: https://research.papercup.com/samples/temporal-control-interspeech-2021

values, and hard to reproduce [21].

Instead of a learnt latent space, using acoustic features directly extracted from reference speech leads to stable and reproducible model estimation as well as ensuring interpretability. Our experimental results (Section 5) demonstrate that the acoustic features are disentangled: they can be independently controlled, unlike the dimensions of the learnt latent space in a comparison model.

2.2. Temporal precision

The use of extracted acoustic features as additional input to a TTS model has recently been explored as a means of prosodic control, especially for F_0 [17, 22, 23]. In [13], the authors demonstrate control over multiple extracted acoustic features, although by learning a *global* embedding that is unable to provide control with temporal precision.

FastSpeech 2 [17] is a non-autoregressive TTS model conditioned on extracted F_0 and energy features, which uses explicit phone durations. However, the features are required per-frame, which is not readily suitable for human-in-the-loop control. FastPitch [23] addresses this challenge by modelling F_0 per-phone, but no longer models energy.

To the best of our knowledge, there is no previous model that provides precise and localised control over all of F_0, energy and duration at an appropriate temporal granularity that balances the competing requirements of i) accounting for unexplained acoustic variation, and ii) intuitive control for a human-in-the-loop.

3. Proposed model and acoustic features

Ctrl-P follows a multi-speaker Tacotron 2 [11, 24] attention-based encoder-decoder architecture, with modifications. A separately-trained WaveRNN vocoder [25] is used to generate a waveform from the mel spectrogram.

Let $p_1, ..., p_N$ denote the sequence of phones to be synthesised, and $\mathbf{y}_1, ..., \mathbf{y}_T$ denote the sequence of frames from the corresponding ground truth mel spectrogram. The proposed modification to the Tacotron 2 architecture consists of concatenating the sequence of encoder outputs $\mathbf{e}_1, ..., \mathbf{e}_N$ with the corresponding, phone-aligned acoustic features $\mathbf{a}_1, ..., \mathbf{a}_N$. In our experiments, the acoustic features are 3-dimensional. The decoder has access to this enhanced representation via the attention mechanism. We now describe how to obtain these acoustic features during training and inference.

3.1. Training

Using forced alignment, each phone p_i aligns to a sequence of frames, $\mathbf{y}_{\alpha(i)}, ..., \mathbf{y}_{\beta(i)}$, from the ground truth mel spectrogram. The forced aligner uses the Kaldi Toolkit [26] with a model trained on the TTS training data.

F_0 is estimated using the RAPT algorithm [27] and root mean square energy using the Librosa library [28]. We take the average of these features per-phone. The duration of phone p_i is represented as the number of frames, $\beta_i - \alpha_i + 1$. For special tokens in the phone sequence, representing word and sentence boundaries, the value of all acoustic features is set to 0.

Each feature is normalised to zero mean and unit standard deviation, per speaker. These three features are then concatenated to form $\mathbf{a}_i$.

In initial experiments, taking the log of feature values [13, 17] did not improve performance. Per-utterance normalisation also degraded performance. Neither were used in the experiments in Section 4.

3.2. Inference

During inference, ground truth acoustic features are not required, but are predicted by the model. The acoustic feature predictor (AFP) takes as input $\mathbf{e}_1, ..., \mathbf{e}_N$ and predicts $\tilde{\mathbf{a}}_1, ..., \tilde{\mathbf{a}}_N$. The AFP consists of two stacked LSTM blocks, each comprising 2-layers of bidirectional LSTMs. The encoder outputs are mapped to a sequence of 64-dimensional hidden states by the first block, then to a sequence of 32-dimensional embeddings by the second block, then through a fully connected layer to a sequence of 16-dimensional embeddings and finally through a $\tanh$ non-linearity, followed by a projection to 3 dimensions to obtain the predicted acoustic features corresponding to each phone in the encoder input: $\tilde{\mathbf{a}}_1, ..., \tilde{\mathbf{a}}_N$.

4. Experimental setup

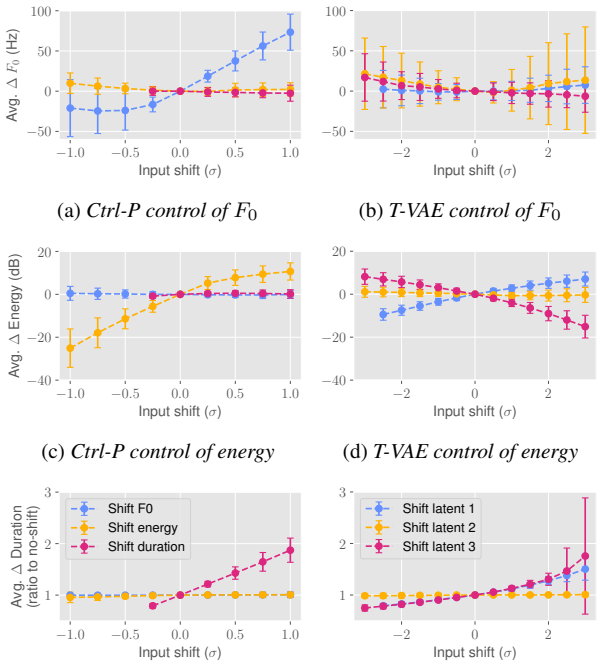
We present experimental results to demonstrate that our proposed model, Ctrl-P, provides acoustic feature control that is more interpretable, disentangled and reproducible than the T-VAE benchmark. We also show that the proposed model generates more natural-sounding speech than the T-VAE or Tacotron-Ref benchmarks described in the next section, and that this naturalness can be further increased by human-in-the-loop control.

4.1. Proposed model and benchmarks

The data set used for all models was a proprietary, multi-speaker, Mexican-Spanish corpus consisting of approximately 38 hours of speech. Of this, $\sim$800 utterances were held-out for validation, with all speakers present in the training set being proportionately represented. We trained our modified Tacotron 2 model for 200k iterations using a weighted sum of gate loss and mel spectrogram reconstruction loss [11]. The model weights were then frozen and the AFP was trained for 400k iterations using an L1 loss between predicted and ground truth acoustic features $\mathbf{a}_1, ..., \mathbf{a}_N$. We used the Adam optimiser in both training phases.

In order to demonstrate the advantage of control using these explicit, extracted features, we compared the performance of our model against a temporal variational auto-encoder (henceforth: T-VAE) similar to the one described in [6] which learns a latent space in an unsupervised fashion. This model uses a secondary attention mechanism to align an input reference mel spectrogram to the encoder outputs and thus produces a sequence of 3-dimensional latents, one per encoder time step (i.e., per phone). A KL divergence loss is applied to the latent space to encourage its distribution to be close to a standard normal. During inference, these latents are predicted from the encoder output using a latent predictor (LP) whose architecture is identical to that of the AFP. The LP is trained in the same way as the AFP, using an L1 loss between the predicted latents and the latents produced from the reference mel spectrogram.

As a benchmark for naturalness, we used another well-established model: Tacotron 2 with a fixed-length global embedding predicted from a reference encoder [5] (henceforth: Tacotron-Ref). Since this model offers no explicit control beyond providing a reference mel spectrogram, we did not include this in our evaluations of controllability.

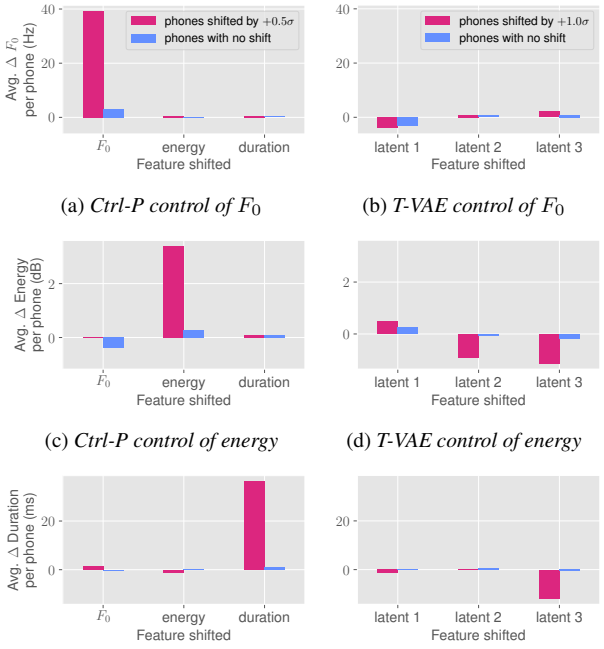


(a) *Ctrl-P control of F_0* (b) *T-VAE control of F_0*

(c) *Ctrl-P control of energy* (d) *T-VAE control of energy*

(e) *Ctrl-P control of duration* (f) *T-VAE control of duration*

Figure 1: *Objective evaluation of disentangled control. x-axis: fraction of the speaker-specific standard deviation by which the feature (or latent) was shifted. Note that T-VAE required much larger changes to obtain comparable acoustic differences. y-axis: resulting change in that feature. Points represent the mean; whiskers denote one standard deviation. Shifting the duration below -0.25σ often generated speech that was too fast for the Kaldi aligner, so these data points are omitted. Results are averages across the validation set.*



(a) *Ctrl-P control of F_0* (b) *T-VAE control of F_0*

(c) *Ctrl-P control of energy* (d) *T-VAE control of energy*

(e) *Ctrl-P control of duration* (f) *T-VAE control of duration*

Figure 2: *Objective evaluation for precision of temporal control. Results are averaged across all 65 validation set utterances generated for a randomly chosen female speaker. Similar behaviour was observed for all speakers.*

4.2. Waveform generation

For the subjective naturalness evaluations, samples were vocoded with model-specific WaveRNN vocoders trained for 3M iterations on the mel spectrograms generated by their respective model for the training set. For the objective controllability evaluations, because subjective quality was not being measured, the large number of waveform samples required was generated using the Griffin-Lim algorithm [29].

5. Results

5.1. Disentangled control

We begin by demonstrating that our proposed model provides disentangled control over each individual acoustic feature. We modified the entire contour of each feature (or latent, in the case of the T-VAE benchmark) dimension by shifting it a fraction of the per-speaker standard deviation for that dimension. The other two dimensions were not shifted.

The average change in utterance-level F_0, energy and duration of the synthesised output was then measured. Figure 1 illustrates the measured changes for each acoustic feature. Figures 1a, 1c and 1e show that the Ctrl-P model is able to make changes only to the specified feature. For example, increasing the F_0 feature results only in an increase in F_0 of the synthesised output. In contrast, T-VAE produces entangled changes (Figures 1b, 1d and 1f). Not only does a change in the value of an individual latent result in a hard-to-interpret change of multiple acoustic properties, those changes can be inconsistent across utterances as seen in the wide standard deviation bands (e.g., changes in F_0 from shifting T-VAE latent 2).

5.2. Temporally-precise control

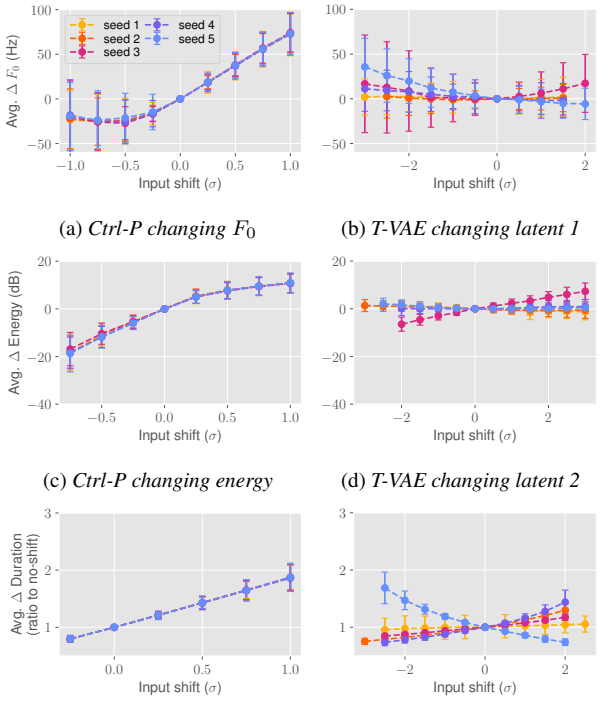
We randomly selected a subset of the stressed vowels within an utterance and shifted each feature (or latent) dimension in turn for that phone, leaving other phones unchanged. Forced alignment was used to label the modified phones in the synthetic output waveform and the change in F_0, energy and duration was measured.

Figure 2 shows that Ctrl-P achieves temporally-precise and disentangled control of only the intended phones, for all three features. In contrast, the temporal region of influence is unclear for T-VAE, with both the modified and unmodified phones undergoing changes in acoustic properties. This might be attributed to the use of an attention mechanism to align latents with phones in T-VAE, in contrast to the hard alignment used by Ctrl-P.

To illustrate the utility of having this fine-grained level of control, we provide samples with varying renditions of the same text that result in semantically distinct utterances.[1]

5.3. Reproducibility

By retraining Ctrl-P and T-VAE from different random seeds, then applying the same analysis as Section 5.1, we are able to quantify the reproducibility of each model. Figure 3 shows that the effect of the T-VAE latent dimensions varies substantially



(a) *Ctrl-P changing* F_0 (b) *T-VAE changing latent 1*

(c) *Ctrl-P changing energy* (d) *T-VAE changing latent 2*

(e) *Ctrl-P changing duration* (f) *T-VAE changing latent 3*

Figure 3: *Objective evaluation of model reproducibility. Shifting each input acoustic feature in Ctrl-P results in predictable changes in the generated speech across random seeds. Shifting each latent dimension in T-VAE results in unpredictable changes.*

across random seeds. There is no such sensitivity for Ctrl-P.

5.4. Naturalness

Figure 4 presents the results from a MUSHRA-like listening test [30] of a randomly chosen selection of 5 validation utterances each from 6 speakers (3 male and 3 female). The samples for Ctrl-P and T-VAE were generated using standard inference (i.e., acoustic features or latents were automatically predicted and not modified). For inference with Tacotron-Ref, the speaker-specific mean embedding estimated from the training set is used.

To create the Ctrl-P (human-in-the-loop) samples, the acoustic features predicted by the AFP were modified by a human, aiming for higher fidelity to the original reference. The uninterpretable and entangled behaviour of the T-VAE latents made such human control impractical for T-VAE. Samples from these 4 models plus a hidden reference (natural speech) and anchor were presented to 50 Spanish-speaking listeners recruited via Amazon Mechanical Turk.

Listeners were asked to rate the naturalness of each sample on a 0-100 scale in intervals of 10, based on how similar to the reference they sound.

We filtered out listeners who failed to identify the hidden reference (by ranking it the highest) more than 50% of the time, leaving 32 valid listeners. Results are presented in Figure 4; all pairs significantly differ in naturalness (two-sided t-test with Holm-Bonferroni correction; $p \leq 0.05$).

We observe that the standard Ctrl-P model is able to generate marginally more natural speech than the T-VAE model. This is despite Ctrl-P being constrained to encoding the supervised acoustic features whilst T-VAE is able to encode *any* information it chooses from the reference mel spectrogram. Both models were found to produce more natural speech than Tacotron-Ref.

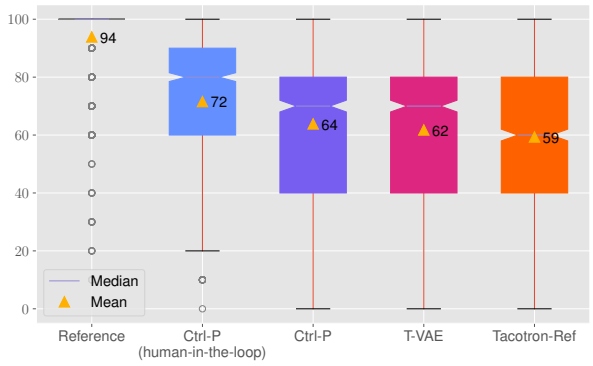


Figure 4: *Results from a MUSHRA-like subjective evaluation of naturalness. Each box spans the 1st to 3rd quartiles (Q1, Q3); whiskers denote the range (capped at* $1.5 \times (Q3 - Q1)$*); outliers are shown as individual points.*

Human-in-the-loop control of the Ctrl-P acoustic features resulted in a further increase in naturalness. The temporal precision offered by the model enabled annotators to make specific, targeted adjustments to the rhythm, intonation and word emphasis within the utterance. Moreover, this modification of the acoustic features did not negatively impact the ability of the matched neural vocoder (trained for Ctrl-P) to generate high-quality waveforms.

6. Conclusions and future work

By modelling F_0, energy and duration explicitly, the proposed model provides interpretable, disentangled, and temporally-precise control over those properties in the generated speech. Model training is reproducible since it is not overly-sensitive to random seed. The chosen feature set could be expanded to include other acoustic correlates of prosody, such as spectral tilt or segmental reduction.

Future work might focus on improving the feature predictions from the AFP by exploring alternate architectures and training routines to obtain improved 'default' prosody for the model, in the absence of, or as a better starting point for, human-in-the-loop control. We also observed a tapering of the influence of control at the extremes of feature values seen during training. Additional research could address this, aiming for a model that is able to generalise *beyond* the range of feature values found in the data.

Our results demonstrate that the model is amenable to human-in-the-loop modifications of the synthetic speech. However, whilst per-phone control over the three principal acoustic correlates of prosody enabled improvements in naturalness to be achieved, it may be preferable to provide more abstract controls such as 'emphasise this word', or 'create rising question intonation'.

Acknowledgements We thank our adviser Mark Gales for feedback on this work.

7. References

[1] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *Proc. ICML*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 3331–3340.

[2] Z. Hodari, O. Watts, and S. King, "Using generative modelling to produce varied intonation for speech synthesis," in *Proc. ISCA SSW*, 2019, pp. 239–244.

[3] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5167–5176.

[4] W. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc ICLR*. OpenReview.net, 2019.

[5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. ICML*, J. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 4693–4702.

[6] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. ICASSP*. IEEE, 2019, pp. 5911–5915.

[7] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc Interspeech*, H. Li, H. M. Meng, B. Ma, E. Chng, and L. Xie, Eds. ISCA, 2014, pp. 1504–1508.

[8] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *arXiv preprint arXiv:1807.11470*, 2018. [Online]. Available: http://arxiv.org/abs/1807.11470

[9] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 905–945, 2010, pMID: 22096264.

[10] S. Sánchez-Mompeán, *The Prosody of Dubbed Speech: Beyond the Character's Words*. Palgrave Macmillan, 02 2020.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.

[12] Z. Zeng, J. Wang, N. Cheng, and J. Xiao, "Prosody learning mechanism for speech synthesis system without text length limit," in *Proc. Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4422–4426.

[13] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," in *Proc. Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4432–4436.

[14] V. Strom, R. A. J. Clark, and S. King, "Expressive prosody for unit-selection speech synthesis," in *Proc. Interspeech*. ISCA, 2006.

[15] T. Masuko, T. Kobayashi, and K. Miyanaga, "A style control technique for hmm-based speech synthesis," in *Proc. Interspeech*. ISCA, 2004.

[16] T. Nose, Y. Kato, and T. Kobayashi, "Style estimation of speech based on multiple regression hidden semi-markov model," in *Proc. Interspeech*. ISCA, 2007, pp. 2285–2288.

[17] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, In Press. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[18] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc ICLR*, 2018.

[19] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *Proc. ICASSP*. IEEE, 2020, pp. 6264–6268.

[20] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," in *Proc. Interspeech*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 4440–4444.

[21] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. ICML*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 4114–4124.

[22] M. Morrison, Z. Jin, J. Salamon, N. J. Bryan, and G. J. Mysore, "Controllable neural prosody synthesis," in *Proc. Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4437–4441.

[23] A. Lancucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *Proc ICASSP*. IEEE, In Press. [Online]. Available: https://arxiv.org/abs/2006.06873

[24] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," in *Proc. Interspeech*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2080–2084.

[25] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 2415–2424.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[27] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, 1995.

[28] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, K. Lee, O. Nieto, J. Mason, D. Ellis, R. Yamamoto, S. Seyfarth, E. Battenberg, V. Moroz, R. Bittner, K. Choi, J. Moore, Z. Wei, S. Hidaka, nullmightybofo, P. Friesch, F.-R. Stöter, D. Hereñú, T. Kim, M. Vollrath, and A. Weiss, "librosa/librosa: 0.7.2," Jan. 2020.

[29] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[30] I. T. U. R. Recommendation, "Bs. 1534-3. method for the subjective assessment of intermediate sound quality (mushra)," *International Telecommunications Union, Geneva*, 2015.